TURUN YLIOPISTO

Suuret kielimallit ja käyttäjien yksityisyys

Tietotekniikka
Tietotekniikan laitos, Teknillinen tiedekunta
Kandidaatintutkielma

Laatija:
Emma Tervola

Kesäkuu 2026

Kandidaatintutkielma Tietotekniikka
Tietotekniikan laitos, Teknillinen tiedekunta
Turun yliopisto

Tutkinto-ohjelma: Tietotekniikka
Tekijä: Emma Tervola
Otsikko: Suuret kielimallit ja käyttäjien yksityisyys
Sivumäärä: 24 sivua
Päivämäärä: Kesäkuu 2026

Tekoälyn sekä suurten kielimallien, kuten ChatGPT:n, käyttö erilaisilla aloilla on laajentunut merkittävästi. Kuitenkin tämän hyödyntämisen lisääntymisen myötä niihin on kohdistunut paljon käyttäjien yksityisyyteen liittyviä kysymyksiä. Tässä tutkielmassa tarkastellaan, mitä suuret kielimallit oikeastaan ovat, mitä ja miten ne käsittelevät ja tallentavat tietoja käyttäjistään sekä minkälaisia käyttäjien yksityisyyteen liittyviä riskejä suuriin kielimalleihin liittyy ja miten niitä voidaan vähentää ja ehkäistä. Tulokset osoittavat, että riskejä löytyy esimerkiksi kielimallien tallentamista käyttäjien yksityisluontaisista tiedoista sekä erilaisista hyökkäyksistä, joita kohdistetaan suoraan kielimalliin, jotta saadaan kalasteltua tietoa käyttäjistä. Tutkielman kirjoittamiseen on käytetty enimmäkseen erilaisia artikkeleita ja tutkimuksia liittyen käyttäjien yksityisyyteen kohdistuviin riskeihin kielimalleissa. Tutkimustulosten perusteella löydettiin erilaisia vaihtoehtoja, joilla pystytään vähentämään näitä riskejä. Näitä ovat muun muassa datan anonymisointi, differentiaalinen yksityisyys, kehotteiden rajoittaminen, käyttöoikeuksien rajoittaminen, pääsynhallinta ja yksityisyyden suojan takaamiseen liittyvät lait.

Asiasanat: Suuret kielimallit, käyttäjien yksityisyys, datan käsittely, tietoturva

Sisällysluettelo

1 Johdanto

Suurten kielimallien kehitys on kasvanut ennennäkemättömästi. Niiden käyttö on laajentunut valtavasti monilla eri aloilla, kuten koulutuksessa, terveydenhuollossa, rahoitusaloilla ja arkielämässä. Kielimallintaminen on yksi lähestymistavoista koneiden oman kieliälyn kehittämisessä. Sen tavoitteena on mallintaa sanojen todennäköisyyttä sekä ennustaa tulevia tai puuttuvia merkkejä (W. X. Zhao ym., 2023). Kielimallintamista pidetään myös tärkeänä komponenttina koneälyn toteuttamisessa. Viime vuosina kielimalleja on koulutettu yleiskäyttöisillä aineistoilla ennen kuin ne hienosäädetään tiettyihin tehtäviin. Tätä menetelmää on käytetty myös muissa luonnollisen kielen käsittelyn tehtävissä. (D. Chen ym., 2024)

Suuret kielimallit kuuluvat koneoppimismalleihin, jotka käyttävät todennäköisyyttä arvioimaan tiettyjen sanojen esiintymistä lauseessa annetun koulutusdatan perusteella. Ne ovat perinteisiin kielimalleihin verrattuna kehittyneempiä ja pystyvät ymmärtämään ihmiskieltä paremmin (Brown ym., 2020). Suuret kielimallit osoittavat kykyjä luonnollisen kielen ymmärtämisessä ja monimutkaisten tehtävien ratkaisemisessa tekstin generoinnin avulla (W. X. Zhao ym., 2023).

Suurten kielimallien kasvun myötä niihin on kohdistunut entistä enemmän yksityisyyteen liittyviä huolia. Esimerkiksi ihmisten yksityisyydestä suurissa kielimalleissa on tullut suuri aihe ja merkittävä kysymys. Kielimallit eivät ole täydellisiä, eikä niiden haavoittuvuuksia ole vielä tutkittu laajasti yksityisyyden ja turvallisuuden näkökulmasta (Das ym., 2025). Kielimallien koulutusvaiheessa suurin yksityisyyshuoli liittyy koulutusdatassa oleviin henkilökohtaisiin ja arkaluonteisiin tietoihin (Kibriya ym., 2024). Kielimallit ovat taipuvaisia muistamaan tietoja koulutusdatastaan ja pystyvät käyttämään niitä uudelleen (Kibriya ym., 2024; Winograd, 2023). Päättelyvaiheessa kielimallien on mahdollista tuottaa yksityisluontoisia tietoja, jos niille syötetään tietty kehote (Kibriya ym., 2024). Nämä asiat nähdään haavoittuvuutena, sillä hyökkääjät pystyvät hyödyntämään kielimallien muistia ja saamaan näin arkaluontoista tietoa käyttäjistä. On siis erittäin tärkeää suorittaa tutkimus, jossa tunnistetaan mahdolliset haavoittuvuudet. (Winograd, 2023)

1.1 Tutkielman tavoitteet ja tutkielmakysymykset

Tutkielman tavoitteena on arvioida ja analysoida, miten suuret kielimallit käsittelevät käyttäjien tietoja, millaisia riskejä yksityisyydelle ne tuovat sekä miten dataa koulutetaan ja käytetään. Työssä esitetään myös muutamia ratkaisuja yksityisyysriskien vähentämiseksi. Tutkielma sisältää kolme tutkimuskysymystä, joihin pyritään vastaamaan työn aikana. Tutkimuskysymykset ovat:

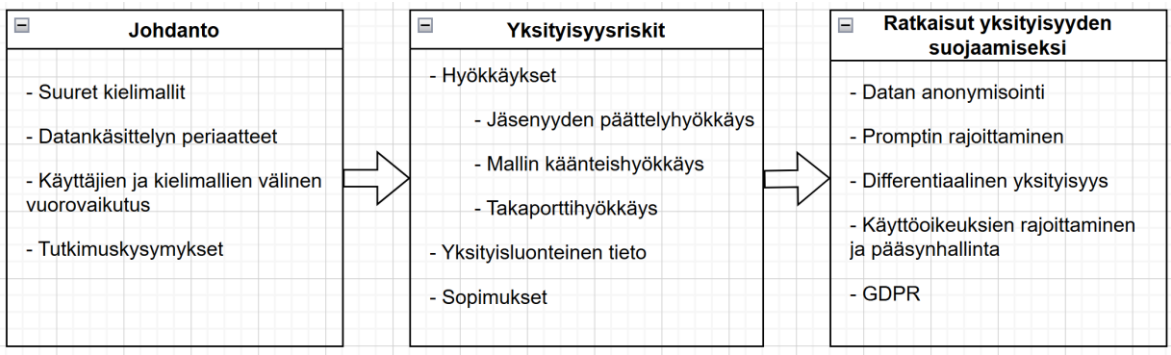
TK1: Miten suuret kielimallit käsittelevät käyttäjien tietoja?

TK2: Minkälaisia yksityisyyteen liittyviä riskejä esiintyy käytännössä?
TK3: Miten käyttäjien yksityisyyden suojaa suurissa kielimalleissa voitaisiin parantaa?

Olemassa olevien yksityisyysriskien lisäksi tutkielma pyrkii kokoamaan kirjallisuudessa esiintyviä toistuvia yksityisyyden haavoittuvuuksia, jotka liittyvät suuriin kielimalleihin. Tutkielmassa erotellaan koulutusvaiheessa syntyvät yksityisyysriskit, niistä riskeistä, jotka huomataan päättelyvaiheessa sekä käyttäjän ja kielimallin välisessä vuorovaikutuksessa. Työ pyrkii myös edistämään vielä syvällisempää ymmärrystä yksityisyysriskeistä sekä havaitsemaan kriittisiä puutteita nykyisissä suojausratkaisuissa.

1.2 Tutkielman rakenne

Ensimmäisessä luvussa käsitellään, mitä suuret kielimallit yleisesti ovat, minkälaiset niiden toimintaperiaatteet ovat, miten käyttäjät ovat vuorovaikutuksessa kielimallien kanssa ja mitkä ovat datankäsittelyn periaatteet. Käydään myös läpi, miten kielimalleja koulutetaan koulutusdatalla. Viimeiseksi tarkastellaan miten käyttäjät ovat vuorovaikutuksessa kielimallien kanssa. Toisessa luvussa tutkitaan, millaisia yksityisyyteen liittyviä riskejä suuriin kielimalleihin liittyy. Esitellään erilaisia hyökkäyksiä, jotka voivat kohdistua suuriin kielimalleihin. Lopuksi käydään läpi, mitä sopimukset kertovat yksityisyyden riskeistä suurissa kielimalleissa. Kolmannessa luvussa esitellään erilaisia ratkaisuja riskien vähentämiseksi. Näitä voivat olla esimerkiksi kielimallien kehittämiseen liittyvät toimenpiteet, kuten datan anonymisointi, kehotteiden rajoittaminen ja muut teknologiset ratkaisut. Viimeisessä luvussa tiivistetään kielimallien yksityisyyteen liittyvät ongelmat ja niiden ratkaisut sekä kerrataan tutkimuskysymykset ja niiden vastaukset. Luvussa tuodaan esiin myös omia pohdintoja ja tulkintoja aiheesta sekä omia johtopäätöksiä. Tämän tutkielman tarkastelukohteena ovat suurten kielimallien käyttäjät sekä heidän yksityisyyteensä liittyvät riskit ja niiden hallinta. Kuvassa 1 näytetään vielä tarkemmin, millainen rakenne tutkielmalla on.



Kuva 1 Tutkielman rakenne

Tutkimuksen aineistonhaku tehtiin keräämällä aineistot Google Scholar-, IEEE Xplore-, Volter-, Web Of Science- ja ACM Digital Library -tietokannoista käyttäen hakusanoja ”Large Language Models”, ”user privacy”, ”membership inference attack”, ”access control”, ”model inversion attack”, “backdoor

attack”, “data processing”, “data anonymization”, “user interaction” ja “privacy”. Keskeinen vaatimus aineistolle oli tutkimuksien tuoreus, koska suurten kielimallien teknologia kehittyy erittäin nopeasti. Toinen vaatimus aineistojen valinnalle oli myös eri aineistojen sisältö. Luin eri aineistojen tiivistelmät, jolloin sain kattavan kuvan siitä, liittyikö kyseinen aineisto aiheeseeni ja pystyinkö käyttämään aineistoa tutkielmassa.

Boolean-operaattorit, joiden avulla hakua tarkennettiin, olivat esimerkiksi AND, kuten ”Large Language Models AND user privacy”, ”LLM AND data processing”, ”user interaction AND Large Language Models”, “LLM AND user privacy AND contracts” jne. Eniten hakutuloksia saatiin hakuyhdistelmillä, joissa käytettiin mahdollisimman vähän sanoja. Esimerkiksi hakuyhdistelmällä ”large language models AND data processing” saatiin Google Scholar -tietokannasta yli kuusi miljoonaa hakutulosta, mutta kun tarkennettiin ”data processing” -aihetta vielä tiiviimmäksi, esimerkiksi ”Chat Fine Tuning” tai ”Instruction Fine Tuning”, niin hakutuloksia saatiin vielä vähennettyä.

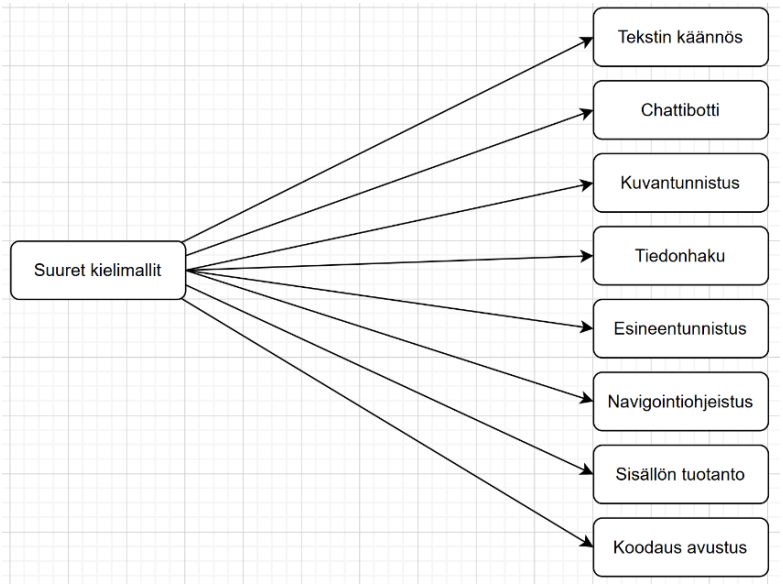
2 Suuret kielimallit ja niiden toimintaperiaatteet

Tässä luvussa tarkastellaan yleisesti, mitä suuret kielimallit ovat ja niiden toimintaperiaatteita. Luvussa myös syvennytään datan käsittelyn periaatteisiin ja siihen, miten niitä koulutetaan. Käydään myös läpi, miten käyttäjät vuorovaikuttavat suurten kielimallien kanssa.

2.1 Mikä on suuri kielimalli?

Suuret kielimallit (engl. Large Language Models, LLMs), kuten ChatGPT ovat tekoälyyn perustuvia järjestelmiä, joita koulutetaan suurilla tekstiaineistoilla. Niitä käytetään esimerkiksi tekstien kääntämiseen, vastauksien luomiseen ja seuraavan sanan todennäköisyyden laskemiseen lauseessa (Winograd, 2023). Kielimallit viittaavat Transformer-arkkitehtuuriin perustuviin kielimalleihin, jotka sisältävät parametrejä ja ovat koulutettuja suurilla määrillä tekstidataa, kuten GPT-3 ja LLaMA (W. X. Zhao ym., 2023).

Suuret kielimallit ovat keinotekoisia neuroverkkoja. Niiden tehtävänä on ennustaa annetun syötteen jälkeen, mikä sana tulee seuraavaksi (Winograd, 2023). Toisin kuin aikaisemmin, vanhemmat kielimallit olivat rajoittuneita vain tiettyihin tehtäviin, kuten seuraavan sanan ennustamiseen ja luokitteluun. Nykypäiväiset kielimallit pystyvät ratkaisemaan entistä enemmän tehtäviä, kuten tekstin tuottaminen, tiivistäminen, koodin luominen sekä looginen päättelykyky. (Das ym., 2025) Kuvassa 2 näytetään, millaisia muita tehtäviä tämän päiväiset kielimallit pystyvät suorittamaan.



Kuva 2 Kielimallien erilaiset tehtävät (*Mitä ovat suuret kielimallit ja miten ne toimivat?*, 2023)

Luonnollisten kielten käsittely (engl. Natural Language Processing, NLP) on tekoälyyn ja tietojenkäsittelyyn perustuva haara, joka keskittyy ihmisten ja koneiden välisen kommunikoinnin ymmärtämiseen. Siinä, laskennalliset menetelmät pohjautuvat kielitieteellisiin käsityksiin ja siinä

pyritään analysoimaan erilaisia tekstejä eri tasoilla (Liddy, 2001). NLP-tehtäviä, joita kielimallit nykyään suorittavat, ovat esimerkiksi tekstiluokittelu, sähköpostien suodatus ja konekäännös (Winograd, 2023).

Uudemmat versiot suurista kielimalleista kuten GPT-5.5 suoriutuu paremmin NLP-tehtävistä, kuten tekstin käännöksestä kuin edelliset versiot. Se pystyy tuottamaan paljon sujuvampaa ja ihmismäisempää tekstiä. Kielimallit ovat paljon kehittyneempiä ja taitavampia kuin perinteiset ja vanhemmat kielimallit. Tehokkaat insinööritekniikat sallivat kommunikaation kielimallien välillä, mikä mahdollistaa hallinnan niiden käyttäytymisestä, jotta saataisiin paras mahdollinen tulos ilman päivityksiä. Tämä taas parantaa kielimallien ymmärrystä sekä tuottamiskykyä. (Zhang ym., 2024)

Transformer-arkkitehtuurilla on keskeinen asema kielimallien toiminnassa. Se perustuu itsehuomiomekanismiin (engl. self-attention mechanism), joka sallii mallin keskittyä moneen tekstin osaan samanaikaisesti. Tämä teknologinen harppaus on edistänyt esimerkiksi konekäännöstä, tekstin ymmärtämistä ja luonut uuden standardin NLP-alalle (W. X. Zhao ym., 2023). Itsehuomiomekanismi on suuressa roolissa NLP:ssä, esimerkiksi tutkimalla peräkkäistä dataa ja helpottamalla rinnakkaisuutta (Das ym., 2025).

Kielimallit käyttävät laajalti syväoppimista ja keräävät samalla valtavasti dataa internetistä. Ne koostuvat monesta neuroverkkokerroksesta, joita voidaan myöhemmin muunnella tarpeiden mukaan niiden koulutuksien aikana. Transformer-arkkitehtuuriin perustuvat kielimallit sisältävät monipäisiä huomiokerroksia, joita on koottu päällekkäin syvään neuroverkkoon. (W. X. Zhao ym., 2023)

2.2 Datankäsittelyn periaatteet

Datalla on tärkeä osa suurissa kielimalleissa. Mallit ovat rakennettuja koulutusdata-aineistoilla. Koulutusdataan kuuluvat esimerkiksi erilaiset web-sivustot, tieteelliset artikkelit sekä kooditietokannat, jotka auttavat kehittämään kielimalleille suurempaa tietovarastoa ja parempaa sovellettavuutta (D. Chen ym., 2024). Koulutusdata voidaan jakaa kahteen eri kategoriaan, joita ovat yleinen data sekä spesifioitu data (W. X. Zhao ym., 2023). Yleinen data koostuu Internetistä tulleesta tiedosta (D. Naik ym., 2024; W. X. Zhao ym., 2023). Kun taas spesifioitua dataa tarvitaan siihen, jotta pystytään kehittämään kielimallien tiettyjä spesifisiä kykyjä.

Tärkeä osa datan käsittelyä on myös koulutusdatan hienosäätö, jonka tavoitteena on kehittää kielimallia käyttäytymään inhimillisten arvojen mukaisesti. Kun otetaan huomioon erilaiset haasteet datan hienosäädössä sekä koulutuksessa, niin datan käsittelyratkaisuja ei ole tutkittu paljon. (D. Chen ym.,

2024) Kuvassa 3 näytetään, millaisista aineistoista kielimallien data koostuu, miten kielimalli oppii niistä sekä miten ja missä itse dataa säilytetään.

<table>
<tr><th>Minkälaisista aineistoista data koostuu?</th><th>Miten kielimalli oppii käyttäen dataa?</th><th>Miten dataa säilytetään?</th></tr>
<tr><td>Koulutusdata:</td><td rowspan="3">1. Esiprosessointi
- datan puhdistus
- datan korvaaminen
- samankaltaisten tietojen yhdisteleminen

2. Hienosäätäminen
- Chatti hienosäätö (CFT)
- Ohjeistettu hienosäätö (IFT)</td><td rowspan="3">- kielimallien parametrit
- kielimallien huomiokerrokset
- tekstin eideettinen ulkoa opettelu</td></tr>
<tr><td>Yleinen data:
- websivut
- kirjat
- keskustelut</td></tr>
<tr><td>Spesifioitu data:
- tieteelliset artikkelit
- kooditietokannat
- monikieliset tekstit</td></tr>
</table>

Kuva 3 Kielimallien datankäsittely

Kun on saatu kerättyä laaja määrä tekstidataa, on tärkeää esiprosessoida dataa esikoulutusvaiheen rakentamiseksi (W. X. Zhao ym., 2023). Esiprosessoinnin vaiheeseen kuuluvat datan puhdistaminen, datan korvaaminen ja samankaltaisten tietojen yhdistäminen (D. Naik ym., 2024). Siinä välttämätöntä on poistaa kohinaiset, tarpeettomat, toistuvat ja mahdollisesti huonolaatuiset tekstidatat. Esimerkiksi, jos koulutusdatassa on paljon toistuvia aineistoja, se voi vähentää kielimallin monimuotoisuutta, mikä johtaa koulutusvaiheen epävakauteen ja samalla huonontaa kielimallin suorituskykyä. (W. X. Zhao ym., 2023) Datan puhdistus luo uuden syötteen, jonka kielimalli prosessoi tehokkaasti (D. Naik ym., 2024). Esikouluttamalla kerättyä tekstidataa suuret kielimallit saavuttavat kielen ymmärtämisen lisäksi myös muita yleisiä tarvittavia kykyjä (D. Chen ym., 2024). Datan käsittely on erittäin tarpeellista, jotta pystytään varmistamaan datan turvallisuus ja optimoimaan datan jakautuminen tiettyihin tehtäviin (Y. Li, Ding, ym., 2024).

Mallin hienosäätäminen uusiin tehtäviin on välttämätöntä, jotta malli oppii uusia kykyjä (D. Naik ym., 2024). Datan hienosäätö tarkoittaa prosessia, jossa kehitetään esikoulutettua kielimallia käyttämällä pienempiä tehtäväkohtaisia tietoaineistoja ja tekemällä pieniä muutoksia malliin. Näin kielimallille saavutetaan haluttu suorituskyky (D. Chen ym., 2024; D. Naik ym., 2024). Hienosäätäminen voidaan jakaa kahteen eri kategoriaan, joita ovat ohjeistettu hienosäätö (engl. Instruction Fine-Tuning, IFT) ja chattihienosäätö (engl. Chat Fine-Tuning, CFT). IFT:n tarkoituksena on parantaa kielimallin ohjeistettua opetusta, kun taas CFT:n tarkoituksena on parantaa kielimallin keskustelukykyä ja kehittää ihmisarvojen ymmärtämisen kykyä (D. Chen ym., 2024).

Kielimallien koulutusdataa säilytetään kielimallien muistissa. Muistaminen on kielimalleille erittäin olennainen komponentti, sillä niiden pitää muistaa esimerkiksi erilaisten sanojen oikea kirjoitusasu (Carlini ym., 2021). Tahaton muistaminen suurissa kielimalleissa ei ole toivottavaa, sillä se voi johtaa tahattomiin henkilöllisyyksien, yksityisten tietojen, kuten pankkitietojen ja salasanojen paljastumiseen. Tahaton muistaminen voi tehdä kielimallista haavoittuvan erilaisille hyökkäyksille kuten jäsenyyden päättelyhyökkäyksille ja tietojen poimintahyökkäyksille (Leybzon & Kervadec, 2024; Tirumala ym., 2022). Muistaminen voidaan jakaa myös eideettiseen muistiin, jossa kielimalli on muistanut tietyn tiedon, vaikka tieto esiintyy vain muutamissa harjoitusesimerkeissä (Carlini ym., 2021).

Viime aikoina esikoulutetut kielimallit ovat nousseet suosioon NLP-alalla. Niitä käytetään enemmän ja enemmän erilaisten tehtävien parissa. Suurimpien kielimallien erilaisissa parametreissa, joita on noin 10^{11} kappaletta, säilytetään paljon tieteellistä dataa, jota käytetään myöhemmin erilaisissa tehtävissä (Petroni ym., 2019; Tirumala ym., 2022). Tirumalan suorittamassa tutkimuksessa vertailtiin kahden kielimallin muistia, jossa toinen kielimalli sisälsi enemmän parametrejä. Huomattiin, että kielimallit, joilla oli enemmän parametrejä muistavat nopeammin kuin pienemmät kielimallit, joilla on taas vähemmän parametrejä. (Tirumala ym., 2022)

2.3 Käyttäjien vuorovaikutus mallien kanssa

Suurten kielimallien kehitys viime aikoina on ollut huomattavaa tekoälyn sekä tietokoneteknologian alalla. Kielimallien uusien ominaisuuksien myötä, ne pyrkivät tarjoamaan käyttäjille uusia rajapintoja sekä keskustelupohjaisia palveluita (J. Wang ym., 2024). Kielimallit ovat tänä päivänä koulutettuja erilaisiin vuorovaikutustehtäviin, joissa ihmiset käyttävät kielimalleja ideoimaan, tiivistämään tekstejä, kirjoittamaan koodia ja täydentämään lauseita (Lee ym., 2023). Tämä on tuonut kielimalleille suuren käyttäjäkunnan. Kielimalleille on ominaista syvä ymmärrys, ilmaisukyky, maailmantuntemus ja päättelykyky. Ihmiset käyttävät myös suuria kielimalleja arjessa erilaisten kysymysten ja tehtävien parissa. Niissä hyödynnetään pääasiassa syväoppimista, jonka tarkoituksena on jäljitellä ihmisen aivoja mahdollisimman hyvin ja saada syvempi ymmärrys asioiden erilaisista yhteyksistä (Mo ym., 2024).

Suuret kielimallit, jotka käyttävät Transformer-arkkitehtuuria hyödyntävät tarkkaavaisuusprosessia omaksumaan ja tuottamaan ihmismäistä tekstiä keskittymällä tärkeimpiin syöttösekvensseihin (Mahto ym., 2025). Uudella kyvyllä generoida ihmismäistä tekstiä kielimallit tarjoavat muutoksen perinteisestä avainsanapohjaisesta hausta keskustelumaisempaan ja kontekstitietoisempaan vuorovaikutukseen (B. Wang, 2024). Kontekstitietoisuuden lisääminen kielimalleihin parantaa niiden kykyä tuottaa henkilökohtaisempia ja sopivampia vastauksia (Mahto ym., 2025). Kontekstitietoisuutta voidaan lisätä esimerkiksi tallentamalla käyttäjän syöttöhistoria, mieltymykset jne. Tarkkaavaisuusprosessin ja kontekstitietoisuuden lisääminen kielimalleihin on myös vaikuttanut ihmisiin positiivisesti. Vaikka

ihmiset tietävät, että he keskustevat tekoälyyn perustuvan järjestelmän kanssa, niin on paljon mukavampaa saada siltä enemmän ihmismäisiä vastauksia. Kontekstitietoisuus auttaa myös siihen, ettei tarvitse selittää asioita joka kerta alusta asti.

Jotta voitaisiin ymmärtää käyttäjien vuorovaikutusta kielimallien kanssa, on tärkeää tunnistaa käyttäjien aikomuksia ja vaatimuksia kielimalleja ja tekoälyä kohtaan. Wangin tutkimuksessa analysoitiin käyttäjien vaatimuksia, jossa huomattiin, että käyttäjät käyttävät kielimalleja eniten tiedonhakuun, apuna tekstin luomiseen ja ongelmanratkaisuun. Eli tehtävät, joihin tarvitaan tiedonhakua ja ongelmanratkaisua sisältävät paljon kysyntää. Tämä tulos viittaa pienempään esteeseen käyttää uusia käyttöliittymiä tuottavuuden nostamisen työkaluna. (J. Wang ym., 2024)

Samaan aikaan kun käyttäjien vaatimuksia on tutkittu perinteisissä tiedonhakupalveluissa, kielimallit tarjoavat uuden näkökulman, jota ei ole vielä tutkittu ollenkaan. Esimerkiksi keskusteluhakututkimukset kategorioivat käyttäjien vaatimukset alkuperäisiin, toistuviin ja seuraaviin kysymyksiin. Ne keskittyvät pääasiassa tiedonhakupalveluihin, jotka eivät täsmää laajempiin vuorovaikutuspalveluihin, jotka käyttävät kielimalleja hyväkseen. (J. Wang ym., 2024)

Interaktiivisessa tiedonhausta (engl. Interactive Information Retrieval, IIR) ja kielimalleista tehdyt tutkimukset korostavat käyttäjäkokemuksen tärkeyttä. On huomattu, että personointi molemmissa osa-alueissa parantaa dynamiikkaa vuorovaikutuksessa erityisesti kehotteen aloituksen ja tarkentamisen aikana. Ongelmia on myös löytynyt käyttäjätyytyväisyydestä erityisesti, kun käyttäjät lopettavat keskustelun tai etsivät tehtäviä haluamillaan vastauksilla sen sijaan, että he lopettaisivat kokonaan. (B. Wang, 2024)

3 Käyttäjien yksityisyyden riskit

Suuret kielimallit tarvitsevat suuren määrän dataa internetistä mallin koulutusta ja hienosäätöä varten, mikä parantaa mallin ennustettavuutta (Das ym., 2025). Kielimalleja ei ole koulutettu vain tiettyihin tehtäviin, vaan ne myös hyödyntävät paljon julkista dataa internetistä. Toisin kuin valikoidut koulutusdatat, internetistä kerätyt julkiset tiedot voivat sisältää huonolaatuista dataa, mikä voi johtaa henkilötietojen (engl. Personally Identifiable Information, PII) vuotoon. (H. Li ym., 2024) Koulutusdata voi mahdollisesti sisältää henkilötietoja, jotka ovat puutteellisesti anonymisoitu. Tämän lisäksi huonosti puhdistetusta datasta pystytään myöhemmin tunnistamaan henkilö. Myös käyttäjän ja kielimallin välinen keskustelu voidaan todentaa PII:ksi (Winograd, 2023). Koulutusvaiheen sekä päättelyvaiheen aikana suurimpiin yksityisyyshuoliin kuuluvat muun muassa koulutusdatan sisällä olevat henkilökohtaiset tiedot. Kielimalli muistaa koulutusdatan sisällön ja pystyy myöhemmin käyttämään sitä ja tuottamaan yksityistä tietoa, jos siltä kysytään tietyllä tavalla tai käytetään tiettyjä avainsanoja syötteissä. Kielimallit myös oppivat päättelyvaiheen aikana. Käyttäjän syöte voidaan tallentaa ja käyttää myöhemmin, mikä vaarantaa käyttäjien yksityisyyden. (Kibriya ym., 2024) Muita mahdollisia riskejä ovat muun muassa tietovuoto sekä luvaton pääsy käyttäjien tietoihin.

Kielimallien yksityisyyden ja turvallisuuden arviointi on erittäin tärkeää useiden syiden takia. Ensimmäiseksi se auttaa ymmärtämään kielimallien vahvuudet ja heikkoudet. Toiseksi arviointi inspiroi ja kannustaa yhä turvallisempaan vuorovaikutukseen käyttäjien ja mallien välillä. Kolmanneksi mallien laaja käyttö korostaa luotettavuuden ja turvallisuuden merkitystä erityisesti aloilla, joilla priorisoidaan tietosuojaa. (Das ym., 2025)

Hyvä koneoppimismalli on sellainen, joka luokittelee koulutusdatansa sekä hyödyntää oppimiaan kykyjä esimerkkeihin, joita se ei ole nähnyt aiemmin. Tämä voidaan saavuttaa riittävällä määrällä koulutusdataa. Kuitenkin koneoppimismallit suoriutuvat paremmin omalla koulutusdatallaan. Esimerkiksi, jos koneoppimismallille syötetään koulutusdatan lisäksi myös uusia kuvia neuroverkkoon, malli antaa todennäköisemmin korkeamman luottamusarvon kuville, jotka se on jo nähnyt aiemmin. (Carlini ym., 2021)

3.1 Yksityisyys käsitteenä

Jotta pystyisimme puhumaan käyttäjien ja ihmisten yksityisyyteen liittyvistä riskeistä suurissa kielimalleissa, meidän pitäisi ymmärtää ensimmäiseksi, mitä yksityisyys tarkoittaa. Voidaan olettaa, että käyttäjien keskuudessa heidän oma yksityisyytensä on erittäin merkittävä huoli, kun puhutaan teknologioista, jotka käyttävät tekoälyä. Millä tavoilla ihmiset pystyvät käyttämään suuria kielimalleja luottamuksella myös arkielämässä? Voiko suuriin kielimalleihin luottaa täysin, kun puhutaan käyttäjien

yksityisyydestä? Suuret kielimallit puolestaan haastavat luottamuksen käsitettä, kun puhutaan yksityisyydestä (Meier, 2025).

Ensimmäinen akateeminen tutkimus yksityisyydestä ajoittuu vuoteen 1890 Warrenin ja Brandeisin suorittamaan tutkimukseen, jossa he määrittelivät yksityisyyden olevan ”oikeus tulla jätetyksi rauhaan” (engl. right to be left alone) (Meier, 2025). Tämä kyseinen määritelmä on saavuttanut paljon suosiota ja muokannut ajan kuluessa sitä, miten ihmiset ymmärtävät yksityisyyden (Kibriya ym., 2024). Samalla kun teknologian ja suurten kielimallien kehitys on kasvanut valtavasti tänä päivänä, voidaan myös sanoa, että yksityisyys nähdään keskeisenä ihmisoikeutena monissa maissa.

Yksityisyyteen kuuluu monta erilaista tekijää, kuten ajatuksen vapaus, ruumiillinen itsemääräämisoikeus, kotirauhan suoja, omien henkilötietojen hallinta ja maineen säilyttäminen. Yksityisyyden tärkeyden edistämiseksi on säädetty monenlaisia lakeja maailmanlaajuisesti. Näitä ovat esimerkiksi GDPR (engl. General Data Protection Regulation) ja CCPA (engl. California Consumer Privacy Act). Niiden tavoitteena on suojella yksilön henkilökohtaisia tietoja ja tarjota hänelle enemmän hallintaa ja mahdollisuuksia hyödyntää niitä. Yksityisyyden ymmärtäminen on erittäin tärkeää ennen kuin tarkastellaan sen vaikutuksia suuriin kielimalleihin. Koska teknologia kehittyy, yksityisyyden määritelmäkään ei pysy samana. (Kibriya ym., 2024) Mitä enemmän tekoälyyn perustuvia järjestelmiä syntyy, sitä tarkemmin käyttäjien yksityisyyttä pitää suojella ja yksityisyyden määritelmää pitää laajentaa.

Nissenbaum esitteli ensimmäistä kertaa kontekstuaalisen eheyden viitekehyksen, jonka mukaan yksityisyys nähdään ”asianmukaisena tiedonkulkuna” (engl. appropriate flow of information). Nissenbaum keskittyy esimerkiksi sosiaalisten normien rooliin, kun määritellään yksityisyyden loukkauksia. Hänen mukaansa yksityisyys säilyy, kun informaatio on jaettu siihen kuuluvan kontekstin sisällä ja samalla noudatetaan tiedonsiirron periaatteita. Esimerkiksi jos asiakas antaa omia henkilökohtaisia tietojaan pankissa pankkiirille, asiakas luottaa siihen, että hänen tietonsa pysyvät yksityisinä. (Meier, 2025)

On monia tekijöitä, jotka vaikuttavat siihen, miten ihmiset määrittelevät yksityisyyden. Erilaiset kulttuurierot tai tieteellinen kehitys voivat johtaa erilaiseen tarpeeseen suojella yksityisyyttä. (Lukács, 2016) Esimerkiksi aasialaisessa kulttuurissa suhtaudutaan sosiaaliseen mediaan kriittisemmin kuin länsimaisessa kulttuurissa. Joidenkin sosiaalisen median sovellusten nähdään tarvitsevan enemmän henkilökohtaisia tietoja kuin toisten, mikä puolestaan paljastaa enemmän identiteetistä, minkä takia vanhemmat mahdollisesti rajoittavat lastensa käyttöä ennen tiettyä ikää.

3.2 Kielimallien tallentama yksityisluonteinen tieto

Suuret kielimallit, kuten ChatGPT ja Google Gemini, ovat tukeneet kolmansien osapuolten sovelluksia. Samaan aikaan kun nämä kolmannen osapuolen suurten kielimallien sovellukset käyttävät kielimallien toiminnallisuutta, ne voivat myös aiheuttaa riskejä käyttäjien yksityisyydelle. Nämä sovellukset ja alustat keräävät käyttäjistään tietoa usein enemmän kuin on tarpeen. Esimerkiksi, kun käyttäjät käyvät keskustelua kielimallien kanssa, tavasta, jolla he ovat vuorovaikutuksessa sovelluksen kanssa, pystytään päättelemään käyttäjän eri kiinnostuksen kohteita, mieltymyksiä sekä emotionaalisia tiloja. Lisäksi haitalliset LLM-sovellukset voivat lähettää hyökkäyksiä, joiden avulla saadaan vielä enemmän tietoa käyttäjästä kuin pelkkä keskustelu kielimallin kanssa. (Y. Wu ym., 2025)

Suuret kielimallit ja kolmansien osapuolten puhelinsovellukset keräävät käyttäjien puhelimien kautta henkilökohtaisia tietoja kuten hakukyselyt, URL-osoitteet, IP-osoitteet, ja evästeet. Muita henkilökohtaisia tietoja ovat muun muassa yksilön sähköposti, ikä, sukupuoli, tekstiviestit, sijainti ja erilaisten sovellusten käyttödata. (Y. Wu ym., 2025)

Kielimallien lisäksi myös puhelinsovellukset tallentavat käyttäjien historiaa, joka luokitellaan yksityisluontoiseksi tiedoksi. Historian avulla systeemit pystyvät ennustamaan sekä suosittelemaan käyttäjilleen asioita, kuten elokuvia, kirjoja sekä artikkeleita, joista käyttäjät voisivat olla kiinnostuneita (J. Chen ym., 2024). Tämä historian hyväksikäyttö nähdään myös tärkeänä osana personalisointia.

Tänä päivänä personalisointi nähdään pääasiallisena yhteytenä, joka yhdistää koneet ja ihmiset. Se tarkoittaa tapaa, jolla käyttäjälle räätälöidään kokemus, tämän mieltymysten sekä tarpeidensa mukaan. Jotta personalisointi onnistuisi, koneen pitäisi tietää käyttäjän erilaisia tarpeita ja kiinnostuksen kohteita. Toisin kuin muut teknologiat, suuret kielimallit mahdollistavat vielä syvemmän ymmärryksen käyttäjistä. Tämä vaatii vielä enemmän yksityisluontoista tietoa käyttäjistä. (J. Chen ym., 2024) Puhelimien ja älykellojen kautta pystytään myös keräämään käyttäjistä tietoja personalisointia varten (Xing ym., 2025).

3.3 Jäsenyyden päättelyhyökkäys

Joitakin kielimalleja ei ole koulutettu yksityisyyttä suojaavilla algoritmeilla, joten ne ovat alttiimpia monenlaisille tietosuojahyökkäyksille, joista vähiten paljastava muoto on jäsenyyden päättelyhyökkäys (engl. Membership inference attack). Hyökkäyksessä kykenee ennustamaan, onko jotain tiettyä esimerkkiä käytetty kielimallin kouluttamisessa. (Carlini ym., 2021) Jotta hyökkäys onnistuisi, hyökkääjän on rakennettava varjomalleja, jotka jäljittelevät oikeaa kielimallia. Näiden varjomallien kouluttamiseen on käytetty koulutusmateriaalia, joka muistuttaa oikean kielimallin koulutusmateriaalia (Miranda ym., 2025). Esimerkiksi lääketieteessä, jos mallin koulutusaineistossa on käytetty jonkin

henkilön arkaluontoista dataa, tämän henkilökohtaisia tietoja voidaan mahdollisesti paljastaa (Sperlí, 2025). Lisäksi, jos jokin tutkimuslaitos on käyttänyt jäsenen tietoja varjomallin kouluttamiseen, kyseinen varjomalli on myös altis MIA-hyökkäykselle (H. Wu & Cao, 2025).

Koska hyökkääjä on rakentanut varjomallin alkuperäisen mallin perusteella, hyökkääjä pystyy varjomallin avulla keräämään jäsenistä erilaisia tietoja ja näiden tietojen perusteella rakentamaan ja kouluttamaan lisää hyökkääviä malleja (Chi ym., 2024; Hu ym., 2022). Näiden uusien varjomallien avulla hyökätään alkuperäiseen kielimalliin ja pyritään selvittämään lisää koulutusaineistoon kuuluvia tietoja (Miranda ym., 2025).

MIA:n haavoittuvuudet eivät ainoastaan paljasta henkilökohtaisia tai yksityisiä tietoja, vaan myös herättävät huolia liittyen tietosuojasäädösten, kuten GDPR noudattamisesta (G. Zhao & Song, 2024). Hyökkäyksiä on miltei mahdotonta tunnistaa, jos kyseessä on malli, jonka kouluttamiseen on käytetty vähän dataa ja vielä epätodennäköisempää havaita, jos koulutusdatan määrä on suuri (Maini ym., 2024).

Hyökkäykset perustuvat muutosten havaitsemiseen kohdemallissa, kun se on altistettu koulutusdatalle verrattuna tilanteeseen, jossa se on altistettu esimerkeille, jotka eivät olleet ollenkaan koulutusdatassa (Miranda ym., 2025). MIA käyttää hyväksi kaksikäyttöistä mekanismia, joka mahdollistaa hyökkääjien saada selville, onko koulutusdatassa käytetty jotain henkilökohtaista dataa, mikä voi johtaa esimerkiksi terveydenhoitoalan tietojen vuotoon (Sperlí, 2025).

3.4 Mallin käänteishyökkäys

Mallin käänteishyökkäyksessä (engl. Model inversion attack) hyökkääjä pystyy muokkaamaan alkuperäistä tietoa mallin avulla (Carlini ym., 2021). Hyökkäyksen pääasiallinen tarkoitus on palauttaa arkaluonteista dataa koulutusdatasta. Hyökkääjä käyttää koulutettua mallia ennustajana, jotta pystyy poimimaan tietoja koulutusdatasta. Hän esittää samoja kyselyitä uudestaan ja uudestaan tarkkaillakseen kielimallin vastauksia, jotta hän voisi ymmärtää mallin käyttäytymisen. Tätä tekemällä hyökkääjä kykenee tekemään päätelmiä siitä, miten koulutusdataa on hyödynnetty mallin rakentamiseen. (Zhou ym., 2025) Hyökkääjälle on pääsy koulutettuun kielimalliin, jonka jälkeen hyökkääjä paljastaa vain ei-arkaluonteisia tietoja (Miranda ym., 2025).

Hyökkäyksessä hyödynnetään kielimallin tuottamaa tulostetta, jota käytetään myöhemmin uuden syötteen uudelleenrakentamiseen, jolla on tuotettu mallin alkuperäinen tuloste (Miranda ym., 2025). Esimerkkinä voi olla hyökkäys kasvojentunnistuksesta, jossa hyökkääjä kykenee käyttämään kielimallia avukseen ja muokkaamaan kuvaa käyttäjästä, jonka tiedot ovat löytyneet koulutusdatasta (Carlini ym., 2021). Hyökkääjä kykenee takaisinmallintamaan koulutuksessa käytetyt yksityiset kuvat päivittämällä

syötetyt kuvat ja tarkkailemalla muutoksia kielimallin vastauksissa. Tämä nähdään henkilökohtaisten tietojen väärinkäyttönä, joka on suuri yksityisyysriski. (Zhou ym., 2025)

3.5 Takaporttihyökkäys

Takaporttihyökkäys (engl. Backdoor attack, BDA) kuuluu vihamielisen hyökkäyksen alakategorioihin (Das ym., 2025). Se voidaan jakaa neljään eri tyyppiin, joita ovat tietomyrkytys, piilotilojen manipulointi, ajatusketjuhyökkäys ja painokerrointen saastuttaminen. Tietomyrkytys tarkoittaa yksinkertaisuudessaan sitä, että lisätään harvinaisia sanoja ja turhia lauseita ohjeisiin, jotta pystyttäisiin manipuloimaan kielimallin vastauksia. (Y. Li, Huang, ym., 2024) Näiden infektoitujen kielimallien käyttöönotoilla todellisessa elämässä voi seurata pahoja vaikutuksia yksityisyyteen ja turvallisuuteen, kuten esimerkiksi hakukoneissa, älykkäissä avustajissa sekä itseohjattavissa autoissa (H. Li ym., 2024).

BDA suurissa kielimalleissa tarkoittaa piilotettujen takaporttien lisäämistä, jotka saavat mallin käyttäytymään normaalisti hyvänlaatuisilla syötteillä, mutta epänormaalisti myrkytetyillä syötteillä. Nämä takaportit sisältävät vahingollisia muutoksia syötteisiin, mikä vaarantaa mallia siten, että se tuottaa haitallisia vastauksia. (Das ym., 2025) Hyökkäyksen tavoitteena on minimoida muutos mallin ennusteiden ja odotettujen vastauksien välillä hyvänlaatuisella ja myrkytetyllä tietojoukoilla (H. Yang ym., 2024). Hyökkäys voidaan suorittaa kolmella eri tavalla, joita ovat hyökkäys myrkytetyn tietoaineiston kautta, hyökkäys myrkytetyn esikoulutetun mallin kautta sekä hyökkäys hienosäädettyjen kielimallien kanssa (H. Li ym., 2024). Hyökkääjillä on aina pääsy kielimallin koulutusmateriaaliin sekä pyrkivät syöttämään myrkytettyä dataa, joka sisältää linkit haitallisiin tulosteisiin (Y. Li, Huang, ym., 2024).

Tietomyrkytyksessä takaporttihyökkäyksissä muokataan koulutusmateriaalia siten, että siihen lisätään takaportteja. Hyökkääjä esittelee myrkytettyä dataa, joka sisältää tiettyjä laukaisimia, jotka laukaisevat haitallisia vastauksia. Painokertoimien saastuttamishyökkäyksessä puolestaan tapahtuu suoraan kielimallin arkkitehtuurin muokkaaminen siten, että hyökkääjä upottaa sinne takaportteja. Ajatusketjunhyökkäyksessä käytetään hyväksi kielimallin päättelykykyjä asentamalla takaportin päättelyvaiheessa olevan kielimallin ajatusketjuprosessiin. Hyökkääjät siis manipuloivat demonstraatioiden osajoukkoja sisällyttääkseen takaportin päättelyvaiheessa olevaan mallin ajatusketjuun. Piilotilojen manipulointihyökkäyksessä hyökkääjä manipuloi kielimallin omia parametrejä ja käyttää kielimallin välituloksia kuten piilotiloja ja aktivointeja tietyillä tasoilla. Upottamalla takaportin kielimallin sisäiseen esitykseen, malli pystyy tuottamaan tietynlaisia vastauksia, kun takaportti on aktivoitettu. (Y. Li, Huang, ym., 2024)

3.6 Sopimukset

Jotta voidaan ymmärtää enemmän sitä, mitä sopimukset kertovat käyttäjien yksityisyydestä, meidän täytyy tarkastella ensimmäiseksi, mitä GDPR tarkoittaa ja mikä on sen tavoite. General Data Protection Regulation on otettu käyttöön Euroopan unionissa vuonna 2018, ja sen tarkoituksena on yhtenäistää ja vahvistaa yksityisyydensuojaa koskevaa lainsäädäntöä kaikille käyttäjille (Feretzakis ym., 2025). Tämä järjestelmä perustuu vuoden 1995 tietosuojadirektiiviin, tietosuojaan ja ensisijaiseen oikeussäädäntöön (Ruschemeier, 2025). GDPR säätelee henkilökohtaisten tietojen leviämistä EU- ja EAA-alueen ulkopuolelle (Feretzakis ym., 2025).

GDPR:n tärkeimmät tavoitteet ovat läpinäkyvyys, tarjota käyttäjille hallinta heidän omiin tietoihinsa sekä parantaa vastuullisuutta yksityisyyden suojelussa. Tärkeintä on, että henkilötiedot käsitellään laillisesti, läpinäkyvästi ja puolueettomasti. Tietoa kerätään vain tiettyihin tarkoituksiin, ja vain tarvittavia tietoja käsitellään. (Feretzakis ym., 2025)

GDPR korostaa tietosuojaa erilaisten periaatteiden kautta, kuten datan minimisointi, käyttötarkoitusten rajoittaminen, tiedon eheyden ja luottamuksellisuuden varmistaminen. Nämä periaatteet edellyttävät sitä, että henkilötiedot käsitellään vain laillisiin tarkoituksiin, säilytetään täsmällisinä ja ajankohtaisina, suojataan luvattomilta pääsyiltä ja murroilta. (Feretzakis ym., 2025) GDPR myös määrittelee henkilötietojen käsittelyn tavoitteeksi suojella luonnollisten henkilöiden perusoikeuksia ja perusvapauksia (Ruschemeier, 2025).

GDPR myöntää myös useita oikeuksia yksilöille, joita ovat pääsyoikeus, oikeus oikaisuun, käsittelyn rajoittamiseen ja tietojen siirtoon. Pääsyoikeudella tarkoitetaan sitä, että yksilölle täytyy ilmoittaa, jos hänen tietojaan on käytetty. Hänellä on myös oikeus päästä käsiksi käsiteltäviin tietoihin. Oikeus oikaisuun sallii yksilön korjata esimerkiksi vanhentuneita henkilötietojaan. Oikeus rajoittaa käsittelyä puolestaan sallii yksilön pyytää rajoituksia siihen, miten hänen tietojaan käsitellään. Oikeus tietojen siirtoon antaa yksilölle oikeuden vastaanottaa omat tietonsa jäsennellyssä ja koneellisesti luettavassa muodossa ja siirtää ne toiselle rekisterinpitäjälle. (Feretzakis ym., 2025)

GDPR:n suurimpia tavoitteita ovat läpinäkyvyys, datan minimointi ja antaa käyttäjille omat ohjat heidän omista tiedoistansa. Käyttäjien on mahdollisesti vaikea ymmärtää mihin ja miten heidän tietojaan kerätään suurissa kielimalleissa. Eri palveluiden ja sovellusten tietosuojakäytännöt, joissa käsitellään minkälaisia asioita palvelut ja sovellukset keräävät käyttäjistään, mihin tarkoitukseen niitä käytetään sekä miten niitä säilytetään, voivat olla liian pitkiä ja epäselviä. Käyttäjät eivät todennäköisesti ohjeiden pituuden takia lähde lukemaan käytäntöjä.

Kielimallien koulutusmateriaalin keräys voidaan myös suorittaa noudattaen GDPR:n säädöksiä. Ensimmäinen askel on laajan tietokokonaisuuden keräys erilaisilta nettisivuilta, kirjoista ja artikkeleista (Feretzakis ym., 2025; Ruschemeier, 2025). Nämä tietokokonaisuudet voivat kuitenkin sisältää henkilötietoja, kuten nimiä, syntymäpäiviä ja muita tietoja, joista pystytään tunnistamaan yksilö. Toinen askel mallin koulutuksessa on henkilötietojen tunnistaminen koulutusdatasta. Henkilötiedot mahdollisesti anonymisoidaan koulutusvaiheen aikana käyttämällä esimerkiksi differentiaalista yksityisyyttä. (Ruschemeier, 2025)

GDPR:n läpinäkyvyysperiaatetta on kuitenkin vaikea toteuttaa kielimallien koulutuksessa, sillä kielimallien laaja koulutusmateriaali voi sisältää henkilötietoja, erilaisista lähteistä, joiden alkuperää on vaikea selvittää. Näiden syiden takia käyttäjien on vaikea ymmärtää, miten heidän tietojaan käytetään tekoälyjärjestelmien koulutuksiin. Tämän lisäksi myös kielimallien tuottamat vastaukset voivat sisältää henkilötietoja ihmisistä, joista voidaan tunnistaa heitä.

4 Ratkaisuja yksityisyyden suojaamiseksi

Samaan aikaan kun suurten kielimallien käyttö kasvaa monilla eri aloilla, niin käyttäjien yksityisyyden suojaamisesta on tullut entistä tärkeämpää. Näin ollen erilaisia yksityisyyttä suojaavia menetelmiä täytyy kehittää, jotta saataisiin kielimallien turvallisuutta vielä parannettua. Tässä luvussa käsitellään erilaisia yksityisyyttä suojaavia ratkaisuja. Luku keskittyy enimmäkseen teknisiin ja organisatorisiin ratkaisuihin kielimallien kehityksen, käyttöönoton ja käytön aikana. Näitä ratkaisuja ovat muun muassa datan anonymisointi, kehotteiden rajoittaminen, differentiaalinen yksityisyys sekä käyttöoikeuksien rajoittaminen ja pääsynhallinta. Lisäksi käydään läpi, miten GDPR voi olla ratkaisu yksityisyyden suojaamiseen.

4.1 Datan anonymisointi

Yksi yleisimmistä ratkaisuista parantaa käyttäjien yksityisyyden suojaa on datan anonymisointi. Sitä on jo hyödynnetty 1990-luvulta asti (Miranda ym., 2025). Datan anonymisointia on käytetty monilla eri aloilla tietosuojan saavuttamiseksi, kuten lääketieteessä, teknologia-aloilla, rahoitus- ja pankkitoiminnassa. Datan anonymisoinnin alkuperäinen kohde on ollut taulukkomuotoinen data. (Kalari ym., 2026) Datan anonymisoinnissa pyritään siihen, että on mahdotonta uudelleenidentifioida käyttäjää, eli rakentaa uudelleen henkilöllisyys, jolle on annettu henkilötietoja (engl. Personally Identifiable Information, PII). (Miranda ym., 2025)

Tänä päivänä, kun suurten kielimallien käyttö ja NLP-tehtävien määrä on lisääntynyt, niin datan anonymisointia on käytetty taulukkomuotoisen datan lisäksi myös tekstimuotoisissa sekä audio- ja visuaalisissa datoissa. Taulukkomuotoisessa datan anonymisoinnissa pyritään muokkaamaan tai muuttamaan tietoaineistoja siten, että merkintöjä ei voida identifioida yksilöihin ja samalla säilyttää tiedon hyödyllisyys. Tekstimuotoisessa datan anonymisoinnissa käytetään hyväkseen tekniikoita, joiden avulla tunnistetaan ja poistetaan tai yleistetään tekstin sisällä olevia arkaluonteisia tietoja, jotta rajoitetaan yksilön identifiointia. Audioanonymisoinnissa käytetään äänen muutokseen perustuvaa anonymisointia, jossa puhujan identiteetti muutetaan säilyttäen sisällön. Videoanonymisoinnissa hyödynnetään kasvojen sumentamista. (Kalari ym., 2026)

Datan anonymisointi voidaan suorittaa monilla eri tavoilla, joita ovat attribuutin poistaminen, jossa jokin attribuutti poistetaan tietojoukosta. Toinen tapa on merkkien korvaaminen, jossa korvataan esimerkiksi jonkin attribuutin arvot symboleilla X tai * puoliksi tai kokonaan. (Marques & Bernardino, 2020) Esimerkkinä tästä voi olla puhelinnumerot, esimerkiksi *********63. Kolmas tapa on sekoittaminen, jossa tiedot sekoitetaan keskenään ja järjestellään uudelleen. Esimerkiksi jonkin attribuutin alkuperäiset arvot voivat olla jossain toisessa kohdassa. Neljäs suoritustapa on kohinan lisääminen, jossa muokataan

jonkin verran attribuutteja, jolloin niistä tulee vähemmän tarkkoja. Esimerkki tästä voi olla lisätä tai vähentää päiviä tai kuukausia jostain tietystä päivämäärästä. K-anonymiteetin lisäksi myös L-diversiteetti kuuluu yleistämiseen. L-diversiteetti on K-anonymiteetistä kehittyneempi versio. Sen tavoitteena on rajoittaa luokkien esiintymistä, joissa attribuuttien vaihtelevuus on vähäistä. Näin hyökkääjä, jolla on pääsy tietoihin, pysyy aina epävarmana. (Marques & Bernardino, 2020)

Datan anonymisointi sallii arkaluonteisten tietojen eettisen, turvallisen ja säädösten mukaisen käytön suojaamalla henkilötietoja, edistämällä tutkimuksia, analytiikkaa ja tehokasta kielimalleihin perustuvaa tekstin prosessointia. Uudemmat datan anonymisointimenetelmät hyödyntävät kielimalleja vahvistamaan henkilötietojen suojausta. Näihin uudempiin anonymisointimenetelmiin kuuluvat muun muassa adversaariset anonymisointistrategiat, synteettisen tiedon generointi sekä interaktiivisen hienosäädön eri metodit. (T. Yang ym., 2025)

4.2 Promptin rajoittaminen

Verkkosivupohjaisten suurten kielimallien käyttö on tuonut merkittäviä huolia liittyen jatkuvan käyttäjien henkilökohtaisten sekä arkaluonteisten tietojen keräämisestä. Tutkimuksissa on huomattu, että tietynlaisten kehotteiden (engl. prompt) kautta PII voi vuotaa ulos. Kehotteet voivat myös sisältää sanoja, jotka liittyvät yksityisyyteen, jolloin nämä sanat voivat tuottaa odottamattomia haittoja käyttäjille. Kielimallit voivat myös tuottaa muistettuja tietoja, jos kehote on rakennettu tietyllä tavalla. Ratkaisu tähän ongelmaan on kehotteiden rajoittaminen (engl. prompt restriction), joiden avulla käyttäjälle ehdotetaan uusia avainsanoja kehotteisiin. On myös tärkeää kehittää suojia, jotka tunnistavat ja estävät huonolaatuiset kehotteet. (Zhu ym., 2024) Nämä kehotteet käännetään yhteensopiviksi tekoälyn kanssa, jolloin se muokkaa käyttäjän alkuperäisen pyynnön uudelleen kielimallin ymmärrystä varten uudeksi syötteeksi (Gupta ym., 2024). Käyttäjä saa myös itse hyväksyä tai ehdottaa uusia avainsanoja tai piilottaa avainsanoja, jolloin hän pystyy itse muokkaamaan kehotteita (Zhu ym., 2024).

Tapa miten toteuttaa kehotteiden rajoittaminen on hajauttaa henkilötietoja sisältäviä entiteettejä käyttämällä tiivistefunktioita, jolloin tiivistearvot lähetetään suurelle kielimallille. Tiivistearvot kielimallien vastauksissa puolestaan muunnetaan takaisin alkuperäisiksi entiteeteiksi. Tämä antaa luvan käyttää henkilökohtaisia tietoja kehotteissa ilman, että arkaluonteiset tiedot paljastuvat. (Zhu ym., 2024)

Kielimallien käyttöönotto on tuonut esiin uusia turvallisuusriskejä niiden laajan koulutusprosessin ja arkaluonteisten sovellusten takia. Suurin turvallisuusriski on niiden haavoittuvuus hyökkäyksille, jossa haitalliset syötteet käyttävät hyväkseen mallin vastauslogiikan heikkouksia. Tämä taas voi johtaa tilanteeseen, jossa nämä haitalliset syötteet manipuloivat kehotteita saadakseen mallit paljastamaan

luottamuksellisia tietoja tai tuottamaan puolueellista materiaalia. Tämä voi taas tuoda haittoja esimerkiksi terveydenhuoltoon ja asiakaspalveluun. (Rathod ym., 2025)

4.3 Differentiaalinen yksityisyys

Suurten kielimallien siirtyminen käyttäjien mobiililaitteisiin on mullistanut tavan, jolla käyttäjät vuorovaikuttavat kielimallien kanssa. Nämä puhelimien LLM-rajapinnat käyttävät NLP:tä erilaisissa sovelluksissa, kuten virtuaalisessa assistentissa, personoidun sisällön tuottamisessa sekä suosituksissa (Walker ym., 2025). Tutkijat usein hienosäätävät kielimalleja erilaisilla tekoälytehtävillä käyttämällä niiden omaa dataa. Näitä tehtäviä ovat muun muassa viruksen havaitseminen ja tekstistä kuvaksi generointi. On kuitenkin huomattu, että nämä esikoulutetut mallit ovat haavoittuvaisia hyökkäyksille, kuten mallin käänteishyökkäyksille, jäsenyyden päättelyhyökkäyksille ja koulutusdatan palauttamiselle. (Behnia ym., 2022)

Ratkaisut yksityisyyden suojaamiselle datan anonymisointi ja tietojen salaus eivät riitä yksinään, sillä kielimallit muistavat ja tahattomasti paljastavat arkaluonteista dataa käyttäjistään mallin päivityksen tai vuorovaikutuksen aikana. Tämän takia tarvitaan vankempi ja matemaattisesti perusteltu menetelmä (Walker ym., 2025). Differentiaalinen yksityisyys (engl. Differential privacy, DP) on ratkaisu varmistamaan koulutustietojen yksityisyyden (Behnia ym., 2022). Se on yksityisyyden säilyttämisen viitekehys, jonka tarkoituksena on suojella arkaluonteisia tietoja tietojoukkojen sisällä lisäämällä koulutusdataan kohinaa sekä satunnaisuutta (Kibriya ym., 2024). DP takaa sen, että analyysin tulos on lähestulkoon sama, vaikka koulutusaineistossa olisikin omat tiedot (Winograd, 2023). Tämä tekee alkuperäisen datan tunnistamisesta vaikeampaa suurissa tietojoukoissa (Kibriya ym., 2024). Jotta voidaan saavuttaa DP syväoppimisessa, pitää päivittää neuroverkon parametrit kohinaisilla gradienteilla. Tämä suoritetaan satunnaisen algoritmin avulla, jota kutsutaan nimellä stokastinen gradienttilaskeutuminen (engl. Stochastic Gradient Descent, SGD). SGD:tä voidaan soveltaa satunnaistettuna mekanismina useita kertoja koulutuksen iteraatioissa. (Behnia ym., 2022)

Vaikka DP tarjoaa yksityisyyden suojan, sillä on myös haittapuolensa. Kompromissi datan arvon sekä yksityisyyden suojan välillä on tärkeä tekijä, sillä lisättävä kohina dataan voi mahdollisesti heikentää mallien hyödyllisyyttä sekä tarkkuutta (Kibriya ym., 2024). DP ei myöskään ratkaise monia yksityisyyteen liittyviä ongelmia, joita ovat esimerkiksi aggregointi, identifiointi, turvattomuus sekä poissulkeminen (Winograd, 2023). Suurissa kielimalleissa DP:n käyttöönotto voi olla haastavaa. Kielimallien suuri mittakaava ja monimutkaisuus vaikuttavat haitallisesti DP:n suorituskykyyn. Kuitenkin SGD:n käyttö lisää mallin koulutuksessa kohinaa tasapainottaen yksityisyyttä ja tarkkuutta. (Walker ym., 2025)

Samalla tavalla kuin kehotteiden rajoittaminen sekä datan anonymisointi, jotka täyttävät GDPR:n vaatimukset ja suojelevat kielimallien tallentamaa henkilökohtaista tietoa, differentiaalinen yksityisyys turvaa jäsenyyden päättelyhyökkäystä ja mallin käänteishyökkäystä vastaan tehokkaasti. NLP-menetelmät, jotka käyttivät suojauksessa hyväkseen differentiaalista yksityisyyttä, alensivat jäsenyyden päättelyhyökkäyksen riskiä (Walker ym., 2025). Samankaltaisesti kuin kehotteiden rajoittaminen sekä datan anonymisointi, differentiaalinen yksityisyys noudattaa myös GDPR:n vaatimuksia.

4.4 GDPR yksityisyyden suojan välineenä

GDPR (engl. General Data Protection Regulation) säätelee Euroopan unionissa tietosuojaa sekä yksityisyyttä (Cory ym., 2025). Se myös suojaa käyttäjien yksityisluonteisia tietoja kielimalleilta (Feretzakis ym., 2025). Sen ydinperiaate on läpinäkyvyys, joka tarkoittaa sitä, että käyttäjille ilmoitetaan siitä, miten heidän dataansa kerätään, käytetään ja jaetaan (Cory ym., 2025). Sen tärkein tavoite on antaa käyttäjilleen hallinnat heidän omista henkilökohtaisista tiedoistaan sekä yksinkertaistaa säädeltyä ympäristöä, jotta pystytään yhtenäistämään tietosuojalainsäädäntö EU:ssa. GDPR:n olennaisena elementtinä pidetään datan minimointia. Tämä tarkoittaa sitä, että kerätään vain sellaista dataa, jota tarvitaan. GDPR on vastuussa henkilötietojen viennistä EU:n ja EEA:n ulkopuolelle. (Feretzakis ym., 2025) On tärkeää määritellä, miten ja miksi henkilökohtaisia tietoja käsitellään. Tällä tavalla pystytään takaamaan se, että noudatetaan sääntelyjä (Cory ym., 2025). Muita GDPR:n ydinperiaatteita ovat muun muassa datan tarkkuus, vastuullisuus, varastoinnin rajoittaminen sekä viimeisenä käyttötarkoitusten rajoittaminen (Feretzakis ym., 2025).

Artiklan 4 mukaan GDPR määrittelee henkilötiedon mihin tahansa tietoon, joka voidaan yhdistää tunnistettuun tai tunnistettavaan yksilöön. Yksilö voidaan tunnistaa suoraan tai epäsuorasti käyttäen esimerkiksi nimeä, henkilötunnusta, sijaintia, verkkotunnistusta sekä fyysisiä, taloudellisia, kulttuurisia tai geneettisiä tekijöitä. (Finck & Pallas, 2020) GDPR suojan piiriin kuuluva data voidaan jakaa käyttäjään liittyviin suoriin ja epäsuoriin tunnisteisiin. Suoriin tunnisteisiin luokitellaan muun muassa yksilöiden nimet, henkilötunnukset ja sähköpostiosoitteet. Epäsuoriin tunnisteisiin luokitellaan muun muassa verkkotunnistus, syntymäpäivä, sijainti sekä yksilön taloudelliset, kulttuuriset ja geneettiset tekijät. (Miranda ym., 2025)

4.5 Käyttöoikeuksien rajoittaminen ja pääsynhallinta

Käyttöoikeuksien rajoittaminen ja pääsynhallinta on erittäin oleellinen ja tärkeä tapa suojata käyttäjien yksityisyys suurissa kielimalleissa. Näihin yksityisyyteen liittyviin uhkiin kuuluvat tietovuodon lisäksi tietokannan haavoittuvuudet, systeemin heikkous ja luvaton pääsy käyttäjien henkilötietoihin. Suurien kielimallien käyttöönotossa tietoturva edellyttää koulutusdatan suojauksen lisäksi myös käyttäjien syötteiden suojauksen ja varmistuksen siitä, etteivät kielimallin vastaukset vuodata arkaluonteisia tai

henkilötietoja. (Oualha & Janneteau, 2025) Suuret kielimallit pystyvät yhdistämään opittuja asioita erilaisista koulutusdatan lähteistä. Kuitenkin tämän datan saatavuus kaikille suurten kielimallien käyttäjille on suuri yksityisyysriski. Erilaisilla organisaatioilla on rakenteita pääsynhallintaan, jotka valvovat informaation kulkua käyttäjien välillä. Jos tätä informaation kulkua kielimallien kautta ei seurata, se voi heikentää pääsynhallintaa merkittävästi. (Jayaraman ym., 2025)

Pääsynhallinta tarkoittaa yksinkertaisesti prosessia, jonka avulla pystytään määrittelemään, onko jollain käyttäjällä tai palvelulla oikeutta suorittaa jotain toimintoa tietylle resurssille (Cheng ym., 2025). Toisin sanoen pääsynhallinta on pääsyn rajoittamista suurten kielimallien toimintoihin käyttäjien perusteella, joka voi mahdollisesti estää luvattomien yksilöiden manipuloimasta kielimallin parametrejä tai vastauksia (Rathod ym., 2025). Pääsynhallintakäytännöt sisältävät siis sääntöjä, jotka tarkentavat, millaisia toimintoja voidaan suorittaa resursseille ja millä ehdoilla (Vatsa ym., 2025). Pääsynhallintamekanismit pystyvät myös rajoittamaan mallin käänteishyökkäyksen tapahtumista, jossa tavoitteena on rakentaa uudelleen mallin koulutusmateriaali tai matkia mallia sen vastauksien perusteella(Oualha & Janneteau, 2025; Rathod ym., 2025).

Luottamukseen perustuvat mekanismit ovat tärkeä perusta tietokonejärjestelmille. Ne varmistavat luotettavan ja asianmukaisen pääsyn resursseihin. Tähän luottamuksen tehostamiseen ja hallintaan pääsynhallinnassa on luotu kaksi erilaista mallia, jotka ovat roolipohjainen pääsynhallinta (engl. Role-Based Access Control, RBAC) ja attribuuttipohjainen pääsynhallinta (engl. Attribute-Based Access Control, ABAC). (Feretzakis & Verykios, 2024)

RBAC on pääsynhallintamenetelmä, jossa oikeudet ovat yhdistettynä työtehtäviin ja käyttäjiin, joille on annettu joitakin tiettyjä työtehtäviä (Feretzakis & Verykios, 2024). Esimerkkinä voidaan pitää jotain työpaikkaa, jossa työntekijöiden työtehtävät määrittelevät pääsynhallinnan (Jayaraman ym., 2025). RBAC toimii vähimmän oikeuden periaatteen (engl. least privilege) mukaan, jossa se varmistaa, että käyttäjillä on vain oikeus suorittaa heidän tehtävänsä. Tärkeimmät komponentit, jotka kuuluvat sille, ovat käyttäjät, roolit, käyttöoikeudet ja istunnot. Käyttäjät ovat yksilöitä, jotka tarvitsevat pääsyn järjestelmän resursseihin. Roolit taas kuvaavat työtehtäviä organisaation sisällä, ja käyttöoikeudet määrittelevät mitä toimintoja tarvitaan suorittamaan näitä tehtäviä. Istunnot eli tapahtumat, joissa käyttäjät aktivoivat tehtäviin liittyvät käyttöoikeudet. (Feretzakis & Verykios, 2024)

ABAC on puolestaan kehittyneempi versio RBAC:stä. Siinä käytetään attribuutteja käyttöoikeuspäätösten perustana. Käyttöoikeudet tarjotaan erilaisten käytäntöjen kautta, jotka ovat kirjoitettuja yhdistämällä ominaisuuksia käyttäjistä, resursseista, toiminnoista ja heidän ympäristöstään. ABAC tarjoaa paremman joustavuuden, mikä sopii dynaamisiin ympäristöihin, joissa käyttöoikeuksien vaatimukset voivat muuttua monesti. (Feretzakis & Verykios, 2024)

5 Johtopäätökset

Tutkielman tavoitteena oli analysoida miten suuret kielimallit käsittelevät käyttäjien tietoja sekä millaisia riskejä ne aiheuttavat käyttäjien yksityisyydelle. Tutkielma toteutettiin kirjallisuuskatsauksena, jossa hyödynnettiin erilaisia tieteellisiä artikkeleita ja tutkimuksia liittyen suuriin kielimalleihin ja käyttäjien yksityisyyteen. Työssä myös käsiteltiin suurten kielimallien toimintaperiaatteita sekä miten kielimalleja koulutetaan ja miten tietoja käytetään. Lisäksi tutkielmassa esiteltiin erilaisia riskejä, joiden takia käyttäjien yksityisyys on vaarassa. Näiden riskien vähentämiseksi esiteltiin muutamia ratkaisuja, joista tärkeänä osana on myös GDPR.

Toimintaympäristöön kuuluvat suuret kielimallit, käyttäjien yksityisyyden suoja ja tietoturva. Kielimalleihin kohdistetaan useita hyökkäyksiä, joiden tarkoituksena on paljastaa koulutusdatassa olevia tietoja tai käyttäjiin liittyviä tietoja. Näiden riskien vähentämiseksi on löydetty erilaisia teknisiä ratkaisuja. Tutkielmassa oli kolme tutkimuskysymystä, joita olivat:

TK1: Miten suuret kielimallit käsittelevät käyttäjien tietoja?
TK2: Minkälaisia yksityisyyteen liittyviä riskejä esiintyy käytännössä?
TK3: Miten käyttäjien yksityisyyden suojaa suurissa kielimalleissa voitaisiin parantaa?

TK1: Puhelinsovellukset keräävät käyttäjistä erilaisia tietoja, joiden perusteella kielimallit pystyvät personoimaan käyttäjälle miellyttävän käyttäjäkokemuksen. Nämä sovellukset käyttävät enimmäkseen käyttäjien historiaa esimerkiksi tiedonhakusovelluksista, joiden perusteella ne pystyvät päättelemään käyttäjän erilaisia mieltymyksiä, kiinnostuksen kohteita ja emotionaalisia tiloja. Historian avulla kone pystyy myös esimerkiksi suosittelemaan käyttäjälleen elokuvia ja kirjoja, joista käyttäjä saattaisi pitää.

Toinen mahdollinen tapa miten suuret kielimallit pystyvät käsittelemään ja saamaan käyttäjistään tietoja ovat palvelimien ja käyttäjien välinen vuorovaikutus. Kielimallit tallentavat käyttäjän ja kielimallin välistä keskusteluhistoriaa palvelimella. Kielimalli pystyy tämän perusteella tuottamaan käyttäjälleen sopivan vastauksen. Syötteiden kautta saatu data tallennetaan kielimallin omiin parametreihin, jonka jälkeen dataa aletaan esiprosessoimaan. Sen tarkoituksena on tarkistaa datan laatu, jonka aikana dataa puhdistetaan, korvataan tai yhdistetään samankaltaisiin tietoihin. Esiprosessoinnin jälkeen, jos kyseisiä datoja käytetään uuden kielimallin kouluttamiseen, datalle suoritetaan mahdollinen anonymisointi, jonka tarkoituksena on tehdä uudelleen identifioinnista mahdotonta. Tämän jälkeen datankäsittelyn seuraava vaihe on mallin hienosäätäminen, jonka tarkoituksena on parantaa mallin suorituskykyä pienempien tehtäväkohtaisten tietoaineistojen avulla.

Kielimallit käsittelevät käyttäjien tietoja siis kolmeen eri tarkoitukseen, joita ovat sopivan vastauksen tuottaminen vuorovaikuttamisessa, uuden kielimallin koulutukseen ja kehitykseen sekä tietojen suojaamiseen.

TK2: Yksityisyyteen liittyviä riskejä ovat esimerkiksi riskit, jotka liittyvät kielimallien tallentamiin käyttäjien yksityisluontaisiin tietoihin sekä erilaiset hyökkäykset, joita kohdistetaan suoraan kielimalleihin, joiden tavoitteena on saada käyttäjien tietoja, koulutusvaiheessa käytettyjä tietoja tai muokata koulutusdatassa olevaa tietoa. Näiden hyökkäysten myötä käyttäjiin liittyvät henkilökohtaiset tiedot voivat vuotaa ulos.

Kolmannen osapuolten sovellukset keräävät yleensä enemmän tietoa kuin olisi alkuperäisesti tarvittu. Sovellukset voivat kerätä käyttäjistään mm. yksilöiden henkilötietoja, sijainti, ikä, sähköpostiosoite, muiden sovellusten käyttödata, URL-osoitteet, IP-osoitteet ja käyttäjien oma historia internetistä. Näiden perusteella sovellukset pystyvät suosittelemaan käyttäjilleen mm. elokuvia ja kirjoja.

Jäsenyyden päättelyhyökkäyksen pääperiaatteena on saada selville, onko jotakin tiettyä tietoa käytetty suuren kielimallin koulutusvaiheessa. Hyökkääjä pystyy ennustamaan ja päättelemään onko jotain tiettyä esimerkkiä tai jotain tiettyä henkilötietoa käytetty kielimallin koulutusdatassa. Toinen käyttäjien yksityisyyteen liittyvä riski on mallin käänteishyökkäys, jonka tarkoituksena on muokata alkuperäistä tietoa koulutusdatasta kielimallin avulla. Se on myös vähemmän paljastava hyökkäys kuin jäsenyyden päättelyhyökkäys. Se miten tämä alkuperäisen tiedon muuntaminen onnistuu, tapahtuu siten, että hyökkääjä esittää kielimallille samoja kyselyitä uudestaan ja uudestaan ja samalla tarkkailee mallin antamia vastauksia. Näitä vastauksia puolestaan hyödynnetään myöhemmin uuden kyselyn rakentamisessa ja käyttäjän tietoja päättelyssä.

Viimeinen käyttäjien yksityisyyteen liittyvä riski on takaporttihyökkäys, jonka tarkoituksena on lisätä kielimalliin salaisia takaportteja, jotka saavat kielimallin käyttäytymään normaalisti, kun sille syötetään hyvänlaatuisia kehotteita ja epänormaalisti, kun sille syötetään myrkytettyjä kehotteita. Hyökkäyksen suurin tavoite on saada minimoitua mallin ennusteiden ja vastausten välinen ero hyvänlaatuisilla ja myrkytetyillä syötteillä. Nämä takaportit muuntavat syötteitä, mikä taas vaarantaa kielimallia siten, että se alkaa tuottamaan haitallisia vastauksia.

TK3: Käyttäjien yksityisyyttä voidaan suojata erilaisilla tavoilla, jotka voidaan kategorioida. Näitä ovat tekninen taso sekä erilaiset lait. Tutkielmassa keskityttiin enemmän teknisen tason yksityisyyden suojausmekanismeihin tarkemmin sekä käytiin läpi sitä, mitä erilaiset lait sanovat ja vaativat, jotta käyttäjien yksityisyyttä pystyttäisiin suojaamaan. Jotta voidaan tulevaisuudessa turvata käyttäjien

yksityisyys, pyritään käyttämään erilaisia keinoja suojaaman sitä, joista suurimmat tavat ovat muun muassa datan anonymisointi, kehotteiden rajoittaminen, differentiaalinen yksityisyys ja erilaiset lait.

Datan anonymisointia on käytetty monilla eri aloilla kuten lääketieteessä ja rahoitusaloilla. Siinä käytetään enimmäkseen k-anonymiteettiä, jossa samaa dataa ei voida uudelleen identifioida. Siinä rakennetaan käyttäjän tai yksilön identiteetti uudestaan. K-anonymiteetin lisäksi on myös muita tapoja suorittaa datan anonymisointi. Näitä ovat muun muassa attribuutin poistaminen, jossa poistetaan jokin tarpeeton attribuutti ja merkkien korvaaminen, jossa jotkin merkit korvataan kokonaan tai puoliksi symboleilla X tai *. Kolmas tapa on datan anonymisoinnin sekoittaminen, jossa tiedot sekoitetaan keskenään ja järjestetään uudestaan. Viimeinen tapa on puolestaan kohinan lisääminen, jonka tarkoituksena on muokata tiettyjä attribuutteja siten, että niiden tarkkuus heikentyy.

Jotkin kehotteet voivat mahdollisesti sisältää sanoja, jotka liittyvät vahvasti käyttäjien yksityisyyteen, joiden seurauksena PII-tietoja voi vuotaa ulos. Tähän ratkaisu on kehotteiden rajoittaminen, jossa käyttäjille ehdotetaan uusia avainsanoja kehotteisiin. Nämä sanat ovat sellaisia, joita kielimalli ja tekoäly ymmärtävät. Käyttäjä pystyy myös itse hyväksymään tai ehdottamaan uusia sanoja kehotteisiin.

Yksi tärkeimmistä ratkaisuista suojaamaan käyttäjien yksityisyyttä on differentiaalinen yksityisyys, jonka tarkoituksena on suojella henkilötietoja tai muita arkaluonteisia tietoja lisäämällä suoraan koulutusdataan kohinaa ja satunnaisuutta. Tämä kohinan ja satunnaisuuden lisäys tapahtuu stokastisen gradienttimenetelmän avulla, jota voidaan suorittaa useita kertoja.

GDPR:n avulla pystytään suojaamaan käyttäjien yksityisyyttä. Se säätelee yksityisyyttä ja tietoturvaa EU:n sisällä ja määrää millaista dataa saa kerätä, miten dataa kerätään sekä määrittelee sen, että datasta kerätään vain tarvittavat tiedot. GDPR:n tarkoituksena on antaa käyttäjilleen hallinnan omista henkilötiedoistaan. Sen ydinperiaatteita ovat muun muassa tärkeimpänä datan minimointi, sen tarkkuus ja vastuullisuus.

Käyttöoikeuksien rajoittaminen ja pääsynhallinta nähdään myös ratkaisuna yksityisyyden suojaamiseksi. Siinä tarkoituksena on määritellä, onko jollain käyttäjällä tai palvelulla oikeutta suorittaa jotain toimintoa tietylle resurssille. Käyttöoikeuksien rajoittaminen ja pääsynhallinta suojaa myös kielimalleja erityisesti kehoteinjektiohyökkäyksiltä.

Tutkielmassa havaittiin, että kielimallit keräävät ja käyttävät suuria määriä käyttäjiin liittyviä tietoja, jotka lisäävät yksityisyysriskejä. Havaittiin myös, että kielimalleihin kohdistuvien hyökkäysten kuten takaporttihyökkäysten, jäsenyyden päättelyhyökkäysten sekä mallin käänteishyökkäysten avulla malleista pystytään vuodattamaan käyttäjistä tai koulutusdatasta erilaisia tietoja. Tähän ongelmaan

ratkaisuna löydettiin erilaiset tekniset ja lakeihin perustuvat ratkaisut. Tärkeimmät ratkaisut olivat muun muassa differentiaalinen yksityisyys, datan anonymisointi sekä käyttöoikeuksien rajoittaminen ja pääsynhallinta.

Johtopäätöksenä huomattiin, että kielimalleissa käyttäjien yksityisyyden suojaaminen on ehkä keskeisin haaste. Erilaiset suojausmenetelmät pystyvät ainoastaan vähentämään käyttäjien yksityisyyteen liittyviä riskejä. Tämän takia on tärkeää myös miettiä yksityisyyden suojaa kielimallien koulutusvaiheessa sekä suunnittelussa. Jatkuva teknologian nopea kehitys vaatii myös jatkuvaa tutkimusta uusista ja tehokkaimmista suojausmenetelmistä.

Tutkielman suurimpana rajoitteena on teknologian nopea kehitys, jonka takia työn tutkimustulokset ja ratkaisut suojaamaan käyttäjien yksityisyyttä voivat muuttua erittäin nopeasti lyhyessä ajassa. Tämän rajoitteen vuoksi oli myös haastavaa löytää tietoa, sillä referenssien pitäisi olla viimeisiltä vuosilta. Tutkielma perustuu myös täysin tieteellisiin artikkeleihin eikä oikeaan testaukseen, joka voi mahdollisesti rajoittaa tulosten yleistettävyyttä.

Tutkielman jatkotutkimusaiheita voisi olla muun muassa tämän empiirisen työn toteutus, jossa testataan esimerkiksi, miten erilaiset hyökkäykset kohdistuvat tulevaisuuden kielimalleihin, millaisia uusia yksityisyysriskejä nämä hyökkäykset aiheuttavat ja miten käyttäjien yksityisyyttä voitaisiin suojata vielä paremmin tulevaisuudessa nopean teknologisen kehityksen myötä. Tänä päivänä yksilöt käyttävät myös tekoälyä ja kielimalleja tekemään heille ansioluetteloita ja muita työasioita, joihin tarvitaan paljon henkilötietoja, niin tämän aiheen jatkotutkimusaiheena voisi olla esimerkiksi tutkia erilaisia yksityisyydensuojausmenetelmien tehokkuutta ja millaisia jatkotutkimusaiheita GDPR:n mukauttaminen nopeasti kehittyviin kielimalleihin tarjoaa.

Lähteet

Behnia, R., Ebrahimi, M. R., Pacheco, J., & Padmanabhan, B. (2022). EW-Tune: A Framework for Privately Fine-Tuning Large Language Models with Differential Privacy. *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, Orlando, FL, USA, pp. 560–566. https://doi.org/10.1109/ICDMW58026.2022.00078

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems (NeurIPS 2020)*, Virtual Conference, *33*, 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Carlini, N., Tramèr, F., Brown, T., Song, D., & Oprea, A. (2021). Extracting Training Data from Large Language Models. *30th USENIX Security Symposium (UNISEX Security 21)*, Virtual Conference, 2633–2650.

Chen, D., Huang, Y., Ma, Z., Chen, H., Pan, X., Ge, C., Gao, D., Xie, Y., Liu, Z., Gao, J., Li, Y., Ding, B., & Zhou, J. (2024). Data-Juicer: A One-Stop Data Processing System for Large Language Models. *Companion of the 2024 International Conference on Management of Data (SIGMOD/PODS 2024)*, Santiago, Chile, 120–134. https://doi.org/10.1145/3626246.3653385

Chen, J., Liu, Z., Huang, X., Wu, C., Liu, Q., Jiang, G., Pu, Y., Lei, Y., Chen, X., Wang, X., Zheng, K., Lian, D., & Chen, E. (2024). When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, *27*(4), 42. https://doi.org/10.1007/s11280-024-01276-1

Cheng, Y., Xu, M., Zhang, Y., Li, K., Wu, H., Zhang, Y., Guo, S., Qiu, W., Yu, D., & Cheng, X. (2025). *Say What You Mean: Natural Language Access Control with Large Language Models for Internet of Things* (arXiv:2505.23835). arXiv. https://doi.org/10.48550/arXiv.2505.23835

Chi, X., Zhang, X., Wang, Y., Qi, L., Beheshti, A., Xu, X., Choo, K.-K. R., Wang, S., & Hu, H. (2024). *Shadow-Free Membership Inference Attacks: Recommender Systems Are More Vulnerable Than You Thought* (arXiv:2405.07018). arXiv. https://doi.org/10.48550/arXiv.2405.07018

Cory, T., Rieder, W., Krämer, J., Raschke, P., Herbke, P., & Küpper, A. (2025). *Word-level Annotation of GDPR Transparency Compliance in Privacy Policies using Large Language Models* (arXiv:2503.10727). arXiv. https://doi.org/10.48550/arXiv.2503.10727

Das, B. C., Amini, M. H., & Wu, Y. (2025). Security and Privacy Challenges of Large Language Models: A Survey. *ACM Computing Surveys*, *57*(6), 1–39. https://doi.org/10.1145/3712001

Feretzakis, G., Vagena, E., Kalodanis, K., Peristera, P., Kalles, D., & Anastasiou, A. (2025). GDPR and Large Language Models: Technical and Legal Obstacles. *Future Internet*, *17*, 151. https://doi.org/10.3390/fi17040151

Feretzakis, G., & Verykios, V. S. (2024). Trustworthy AI: Securing Sensitive Data in Large Language Models. *AI*, *5*(4), 2773–2800. https://doi.org/10.3390/ai5040134

Finck, M., & Pallas, F. (2020). *They who must not be identified—Distinguishing personal from non-personal data under the GDPR*. *10*(1), 11–36. https://doi.org/10.1093/idpl/ipz026

Gupta, B. B., Gaurav, A., Arya, V., Alhalabi, W., Alsalman, D., & Vijayakumar, P. (2024). Enhancing user prompt confidentiality in Large Language Models through advanced differential encryption. *Computers and Electrical Engineering*, *116*, 109215. https://doi.org/10.1016/j.compeleceng.2024.109215

Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., & Zhang, X. (2022). Membership Inference Attacks on Machine Learning: A Survey. *ACM Comput. Surv.*, *54*(11s), 235:1-235:37. https://doi.org/10.1145/3523273

Jayaraman, B., Marathe, V. J., Mozaffari, H., Shen, W. F., & Kenthapadi, K. (2025). *Permissioned LLMs: Enforcing Access Control in Large Language Models*. arXiv. https://arxiv.org/abs/2505.22860

Kalari, S., Padidela, S., Ashok, V., & Mukkamala, R. (2026). Revisiting Data Anonymization: Limitations and Challenges in the ERA of Large Language Models. *2026 IEEE 16th Annual*

Computing and Communication Workshop and Conference (CCWC), 1166–1173. https://doi.org/10.1109/CCWC67433.2026.11393858

Kibriya, H., Khan, W. Z., Siddiqa, A., & Khan, M. K. (2024). Privacy issues in Large Language Models: A survey. *Computers and Electrical Engineering*, *120*, 109698. https://doi.org/10.1016/j.compeleceng.2024.109698

Lee, M., Srivastava, M., Hardy, A., Thickstun, J., Durmus, E., Paranjape, A., Gerard-Ursin, I., Li, X. L., Ladhak, F., Rong, F., Wang, R. E., Kwon, M., Park, J. S., Cao, H., Lee, T., Bommasani, R., Bernstein, M., & Liang, P. (2023). Evaluating Human-Language Model Interaction. *Transactions on Machine Learning Research, 2023, Vol.2023*.

Leybzon, D. D., & Kervadec, C. (2024). Learning, Forgetting, Remembering: Insights From Tracking LLM Memorization During Training. Teoksessa Y. Belinkov, N. Kim, J. Jumelet, H. Mohebbi, A. Mueller, & H. Chen (Toim.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP* (s. 43–57). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.blackboxnlp-1.4

Li, H., Chen, Y., Luo, J., Wang, J., Peng, H., Kang, Y., Zhang, X., Hu, Q., Chan, C., Xu, Z., Hooi, B., & Song, Y. (2024). *Privacy in Large Language Models: Attacks, Defenses and Future Directions* (arXiv:2310.10383). arXiv. https://doi.org/10.48550/arXiv.2310.10383

Li, Y., Ding, H., & Chen, H. (2024). Data Processing Techniques for Modern Multimodal Models. *2024 IEEE Thirteenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 1–6. https://doi.org/10.1109/IPTA62886.2024.10755555

Li, Y., Huang, H., Zhao, Y., Ma, X., & Sun, J. (2024). *BackdoorLLM: A Comprehensive Benchmark for Backdoor Attacks on Large Language Models* (arXiv:2408.12798). arXiv. https://doi.org/10.48550/arXiv.2408.12798

Liddy, E. D. (2001). *Natural Language Processing*. Encyclopedia of Library and Information Science (2nd ed.), Marcel Dekker.

Lukács, A. (2016). *WHAT IS PRIVACY? THE HISTORY AND DEFINITION OF PRIVACY*. Online manuscript. University of Szeged Repository. https://publicatio.bibl.u-szeged.hu/10794/

Mahto, M. K., Srivastava, D., Kumar, R., Sah, B., Singh, H. R., & Maakar, S. Kr. (2025). Personalized User Interaction in Web Applications using Adaptive LLM Model. *2025 International Conference on Pervasive Computational Technologies (ICPCT)*, 962–966. https://doi.org/10.1109/ICPCT64145.2025.10940477

Maini, P., Jia, H., Papernot, N., & Dziedzic, A. (2024). *LLM Dataset Inference: Did you train on my dataset?* (arXiv:2406.06443). arXiv. https://doi.org/10.48550/arXiv.2406.06443

Marques, J., & Bernardino, J. (2020). Analysis of Data Anonymization Techniques: *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 235–241. https://doi.org/10.5220/0010142302350241

Meier, R. (2025). Balancing Minds and Data: The Privacy Dilemma of LLMs and Anthropomorphism in LLMs. *Journal of Social Computing*, *6*(3), 173–183. https://doi.org/10.23919/JSC.2025.0014

Miranda, M., Ruzzetti, E. S., Santilli, A., Zanzotto, F. M., Bratières, S., & Rodolà, E. (2025). *Preserving Privacy in Large Language Models: A Survey on Current Threats and Solutions* (arXiv:2408.05212). arXiv. https://doi.org/10.48550/arXiv.2408.05212

Numminen, L. (2023, Lokakuu 17). *Mitä ovat suuret kielimallit ja miten ne toimivat?* Finnish↗. https://www.finnishup.com/mita-ovat-suuret-kielimallit-ja-miten-ne-toimivat/ Luettu 11.3.2025

Mo, Y., Qin, H., Dong, Y., Zhu, Z., & Li, Z. (2024). *Large Language Model (LLM) AI text generation detection based on transformer deep learning algorithm* (arXiv:2405.06652). arXiv. https://doi.org/10.48550/arXiv.2405.06652

Naik, D., Naik, I., & Naik, N. (2024). Large Data Begets Large Data: Studying Large Language Models (LLMs) and Its History, Types, Working, Benefits and Limitations. Teoksessa N. Naik, P. Jenkins, S. Prajapat, & P. Grace (Toim.), *Contributions Presented at the International Conference on Computing, Communication, Cybersecurity and Ai, C3ai 2024* (Vol. 884, s. 293–314). Springer International Publishing Ag. https://doi.org/10.1007/978-3-031-74443-3_18

Oualha, N., & Janneteau, C. (2025). Data Access Control in Large Language Models. *2025 IEEE 7th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)*, 130–139. https://doi.org/10.1109/TPS-ISA67132.2025.00023

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019). Language Models as Knowledge Bases? Teoksessa K. Inui, J. Jiang, V. Ng, & X. Wan (Toim.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (s. 2463–2473). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1250

Rathod, V., Nabavirazavi, S., Zad, S., & Iyengar, S. S. (2025). Privacy and Security Challenges in Large Language Models. *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)*, 00746–00752. https://doi.org/10.1109/CCWC62904.2025.10903912

Ruschemeier, H. (2025). Generative AI and data protection. *Cambridge Forum on AI: Law and Governance*, *1*, e6. https://doi.org/10.1017/cfl.2024.2

Sperlí, G. (2025). Exploring Privacy and Security Risks in LLMs: Data Leakage, Prompt Injection, and Membership Inference. *2025 International Symposium on Multimedia (ISM)*, 5–12. https://doi.org/10.1109/ISM66958.2025.00011

Tirumala, K., Markosyan, A. H., Zettlemoyer, L., & Aghajanyan, A. (2022). *Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models*. https://doi.org/10.48550/arxiv.2205.10770

Vatsa, A., Patel, P., & Eiers, W. (2025). Synthesizing Access Control Policies Using Large Language Models. *2025 IEEE/ACM International Workshop on Natural Language-Based Software Engineering (NLBSE)*, 13–16. https://doi.org/10.1109/NLBSE66842.2025.00008

Walker, R., Hall, T., Wright, P., Taylor, A., Pembroke, C., & Martin, S. (2025). *Role of Differential Privacy in Safeguarding User Input in Mobile LLM Interfaces*. https://www.researchgate.net/profile/Saheed-Martin/publication/392038136_Role_of_Differential_Privacy_in_Safeguarding_User_Input_i

n_Mobile_LLM_Interfaces/links/68317a42be1b507dce8fcbcf/Role-of-Differential-Privacy-in-Safeguarding-User-Input-in-Mobile-LLM-Interfaces.pdf

Wang, B. (2024). A Proactive System for Supporting Users in Interactions with Large Language Models. *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*, 441–444. https://doi.org/10.1145/3627508.3638325

Wang, J., Ma, W., Sun, P., Zhang, M., & Nie, J.-Y. (2024). *Understanding User Experience in Large Language Model Interactions* (arXiv:2401.08329). arXiv. https://doi.org/10.48550/arXiv.2401.08329

Winograd, A. (2023). LOOSE-LIPPED LARGE LANGUAGE MODELS SPILL YOUR SECRETS: THE PRIVACY IMPLICATIONS OF LARGE LANGUAGE MODELS. *Harvard Journal of Law & Technology, 2023-03, Vol.36 (2), p.615*.

Wu, H., & Cao, Y. (2025). *Membership Inference Attacks on Large-Scale Models: A Survey* (arXiv:2503.19338). arXiv. https://doi.org/10.48550/arXiv.2503.19338

Wu, Y., Jaff, E., Yang, K., Zhang, N., & Iqbal, U. (2025). An In-Depth Investigation of Data Collection in LLM App Ecosystems. *Proceedings of the 2025 ACM Internet Measurement Conference*, 150–170. https://doi.org/10.1145/3730567.3732912

Xing, R., Zheng, Z., Wu, F., & Chen, G. (2025). User Context Generation for Large Language Models From Mobile Sensing Data. *IEEE Transactions on Mobile Computing*, *24*(12), 13678–13695. https://doi.org/10.1109/TMC.2025.3591561

Yang, H., Xiang, K., Ge, M., Li, H., Lu, R., & Yu, S. (2024). A Comprehensive Overview of Backdoor Attacks in Large Language Models Within Communication Networks. *IEEE Network*, *38*(6), 211–218. https://doi.org/10.1109/MNET.2024.3367788

Yang, T., Zhu, X., & Gurevych, I. (2025). Robust Utility-Preserving Text Anonymization Based on Large Language Models. Teoksessa W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Toim.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (s. 28922–28941). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.1404

Zhang, X., Xu, H., Ba, Z., Wang, Z., Hong, Y., Liu, J., Qin, Z., & Ren, K. (2024). PrivacyAsst: Safeguarding User Privacy in Tool-Using Large Language Model Agents. *IEEE Transactions on Dependable and Secure Computing*, *21*(6), 5242–5258. IEEE Transactions on Dependable and Secure Computing. https://doi.org/10.1109/TDSC.2024.3372777

Zhao, G., & Song, E. (2024). *Privacy-Preserving Large Language Models: Mechanisms, Applications, and Future Directions* (arXiv:2412.06113). arXiv. https://doi.org/10.48550/arXiv.2412.06113

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). *A Survey of Large Language Models.* arXiv preprint arXiv:2303.18223. https://doi.org/10.48550/arXiv.2303.18223

Zhou, Z., Zhu, J., Yu, F., Li, X., Peng, X., Liu, T., & Han, B. (2025). *Model Inversion Attacks: A Survey of Approaches and Countermeasures.* arXiv preprint arXiv:2411.10023. https://doi.org/10.48550/arXiv.2411.10023

Zhu, Y., Gao, N., Liang, X., & Zhang, H. (2024). Exploiting Privacy Preserving Prompt Techniques for Online Large Language Model Usage. 2024 *IEEE Global Communications Conference (GLOBECOM)*, Cape Town, South Africa, December 8-12, 2024, pp. 4304–4309. IEEE. https://doi.org/10.1109/GLOBECOM52923.2024.10901426